APPENDIX 1

Abstract

We disclose a method and apparatus for recognizing sound, music, and other similar signals. The disclosed invention is capable of recognizing an exogenous sound signal that is a rendition of a known recording indexed in a database. The exogenous sound signal may be subjected to distortion and interference, including background noise, talking voices, compression artifacts, band-limited filtering, transmission dropouts, time warping, and other linear and nonlinear corruptions of the original signal. The algorithm is capable of identifying the corresponding original recording from a large database of recordings in time proportional to the logarithm of the number of entries in the database. Given sufficient computational power the system can perform the identification in nearly realtime, i.e. as the sound is being sampled, with a small lag.

Database construction

The sound database may consist of any collection of recordings, such as speech, music, advertisements, or sonar signatures.

Indexing

In order to index the sound database, each recording in the library is subjected to landmarking and fingerprinting analysis to generate an index set for each item. Each recording in the database has a unique index, sound_ID.

Landmarking

Each sound recording is landmarked using methods to find distinctive and reproducible locations within the sound recording. The ideal landmarking algorithm will be able to mark the same points within a sound recording despite the presence of noise and other linear and nonlinear distortion. The landmarking method is conceptually independent of the fingerprinting process, but may be chosen to optimize performance of the latter. Landmarking results in a list of timepoints {landmark$_k$} within the sound recording at which fingerprints should be calculated. A good landmarking scheme marks about 5-10 landmarks per second of sound recording, of course depending on the amount of activity within the sound recording.

Power Norms

A simple landmarking technique is to calculate the instantaneous power at every timepoint and to select local maxima. One way of doing this is to calculate the envelope by rectifying and filtering the waveform directly. Another way is to calculate the Hilbert transform (quadrature) of the signal and use the sum of the magnitudes squared of the Hilbert transform and the original signal.

Spectral Lp Norms

The power norm method of landmarking is especially good for finding transients in the sound signal. The power norm is actually a special case of the more general Spectral Lp Norm, where p=2. The general Spectral Lp Norm is calculated at each time along the sound signal by calculating the spectrum, for example via a Hanning-windowed

Fast Fourier Transform (FFT). The Lp norm for that time slice is then calculated as the sum of the p-th power of the absolute values of the spectral components, optionally followed by taking the p-th root. As before, the landmarks are chosen as the local maxima of the resulting values over time.

Multislice landmarks

Multi-slice landmarks may be calculated by taking the sum of p-th powers of absolute values of spectral components over multiple timeslices instead of a single slice. Finding the local maxima of this extended sum allows optimization of placement of the multislice fingerprints, described below.

Fingerprinting

The algorithm computes a fingerprint at each landmark timepoint in the recording. The fingerprint is generally a value or set of values that summarize a set of features in the recording near the timepoint. In our implementation the fingerprint is a single numerical value that is a hashed function of multiple features.

The following are a few possible fingerprint categories.

Salient Spectral Fingerprints

In the neighborhood of each landmark timepoint a frequency analysis is performed to extract the top several spectral peaks. A simple such fingerprint value is just the single frequency value of the strongest spectral peak. The use of such a simple peak

resulted in surprisingly good recognition in the presence of noise, but resulted in many false positive matches due to the non-uniqueness of such a simple scheme. Using fingerprints consisting of the two or three strongest spectral peaks resulted in fewer false positives, but in some cases created a susceptibility to noise if the second-strongest spectral peak was not sufficiently strong enough to distinguish it from its competitors in the presence of noise – the calculated fingerprint value would not be sufficiently stable. Despite this, the performance of this case was also good.

Multislice Fingerprints

In order to take advantage of the time-evolution of many sounds a set of timeslices is determined by adding a set of offsets to a landmark timepoint. At each resulting timeslice a Salient Spectral Fingerprint is calculated. The resulting set of fingerprint information is then combined to form one multitone fingerprint. Each such fingerprint is much more unique than the single-time salient spectral fingerprint since it tracks temporal evolution, resulting in fewer false matches. Our experiments indicate that using two or three timeslices along with the single strongest spectral peak in each timeslice results in very good performance, even in the presence of significant noise.

LPC Coefficients

In addition to finding the strongest spectral components, there are other spectral features that can be extracted and used as fingerprints. LPC analysis extracts the linearly predictable features of a signal, such as spectral peaks, as well as spectral shape. LPC coefficients of waveform slices anchored at landmark positions can be used as

fingerprints by hashing the quantized LPC coefficients into an index value. LPC is well-known in the art of digital signal processing.

Cepstral Coefficients

Cepstral coefficents are useful as a measure of periodicity and may beused to characterize signals that are harmonic, such as voices or many musical instruments. A number of cepstral coefficients may hashed together into an indes and used as a fingerprint. Cepstral analysis is well-known in the art of digital signal processing.

Index Set

The resulting index set for a given sound recording is a list of pairs (fingerprint, landmark) of analyzed values. Since the index set is composed simply of pairs of values, it is possible to use multiple landmarking and fingerprinting schemes simultaneously. For example, one landmarking/fingerprinting scheme may be good at detecting unique tonal patterns, but poor at identifying percussion, whereas a different algorithm may have the opposite attributes. Use of multiple landmarking/fingerprinting strategies results in a more robust and richer range of recognition performance. Different fingerprinting techniques may be used together by reserving certain ranges of fingerprint values for certain kinds of fingerprints. For example, in a 32-bit fingerprint value, the first 3 bits may be used to specify which of 8 fingerprinting schemes the following 29 bits are encoding.

Searchable Database

Once the index sets have been processed for each sound recording in the database, a searchable database is constructed in such a way as to allow fast (log-time) searching. This is accomplished by constructing a list of triplets (fingerprint, landmark, sound_ID), obtained by appending the corresponding sound_ID to each doublet from each index set. All such triplets for all sound recordings are collected into a large index list. In order to optimze the search process, the list of triplets is then sorted according to the fingerprint. Fast sorting algorithms are well-known in the art and extensively discussed in D.E. Knuth, "The Art of Computer Programming, Volume 3: Sorting and Searching," hereby incorporated by reference. High-performance sorting algorithms can sort the list in N log(N) time, where N is the number of entries in the list. Once this list is sorted it is further processed by segmenting it such that each unique fingerprint in the list is collected into a new master index list. Each entry in this master index list contains a fingerprint value and a pointer to a list of (landmark, sound_ID) pairs. Rearranging the index list in this way is optional, but saves memory since each fingerprint value only appears once. It also speeds up the database search since the effective number of entries in the list is greatly reduced to a list of unique values.

Alternatively, the master index list could also be constructed by inserting each triplet into a B-tree with non-unique fingerprints hanging off a linked list. Other possibilites exist for constructing the master index list. The master index list is preferably held in system memory, such as DRAM, for fast access.

Recognition system

Once the master index list has been built it is possible to perform sound recognition over the database.

Sound source

Exogenous sound is provided from any number of analog or digital sources, such as a stereo system, television, Compact Disc player, radio broadcast, telephone, mobile phone, internet stream, or computer file. The sounds may be realtime or offline. They may be from any kind of environment, such as a disco, pub, submarine, answering machine, sound file, stereo, radio broadcast, or tape recorder. Noise may be present in the sound signal, for example in the form of background noise,
talking voices, etc.

Input to the recognition system

The sound stream is then captured into the recognition system either in realtime or presented offline, as with a sound file. Realtime sounds may be sampled digitally and sent to the system by a sampling device such as a microphone, or be stored in a storage device such as an answering machine, computer file, tape recorder, telephone, mobile phone, radio, etc. The sound signal may be subjected to further degradation due to limitations of the channel or sound capture device. Sounds may also be sent to the recognition system via an internet stream, FTP, or as a file attachment to email.

Preprocessing

Once the sound signal has been converted into digital form it is processed for recognition. As with the construction of the master index list, landmarks and fingerprints are calculated. In fact, it is advisable to use the very same code that was used for processing the sound recording library to do the landmarking and fingerprinting of the exogenous sound input. The resulting index set for exogenous sound sample is also a list of pairs (fingerprint,landmark) of analyzed values.

Searching

Searching is carried out as follows: each fingerprint/landmark pair ($fingerprint_k$,$landmark_k$) in the resulting input sound's index set is processed by searching for fingerprintk in the master index list. Fast searching algorithms on an ordered list are well-known in the art and extensively discussed in Knuth, Volume 3 (ibid), incorporated by reference. If $fingerprint_k$ is found then the corresponding list of matching ($landmark^*_j$, $sound\_ID_j$) pairs having the same fingerprint is copied and augmented with landmarkk to form a set of triplets of the form($landmark_k$, $landmark^*_j$,$sound\_ID_j$). This process is repeated for all k ranging over the input sound's index set, with the all the resulting triplets being collected into a large candidate list.

After the candidate list is compiled it is further processed by segmenting according to sound_ID. A convenient way of doing this is to sort the candidate list according to sound_ID, or by insertion into a B-tree. The result of this is a list of candidate sound_IDs, each of which having a scatter list of pairs of landmark timepoints, ($landmark_k$,$landmark^*_j$) with the sound_ID stripped off.

Scanning

The scatter list for each sound_ID is analyzed to determine whether it is a likely match.

Thresholding

One way to eliminate a large number of candidates is to toss out those having a small scatter list. Clearly, those having only 1 entry in their scatter lists cannot be matched.

Alignment

A key insight into the matching process is that the time evolution in matching sounds must follow a linear correspondence, assuming that the timebases on both sides are steady. This is almost always true unless thesound on one side has been nonlinearly warped intentionally or subject to defective playback equipment such as a tape deck with a warbling speed problem. Thus, the matching fingerprints yielding correct landmark pairs ($landmark_n$,$landmark^*_n$) in the scatter list of a given sound_ID must have a linear correspondence of the form

$$landmark^*_n = m*landmark_n + offset$$

where m is the slope, and should be near 1, $landmark_n$ is the corresponding timepoint within the exogenous sound signal,$landmark^*_n$ is the corresponding timepoint within the

library sound recording indexed by sound_ID, and offset is the time offset into the library sound recording corresponding to the beginning of the exogenous sound signal.

This relationship ties together the true landmark/fingerprint correspondences between the exogenous sound signal and the correct library sound recording with high probability, and excludes outlier landmark pairs. Thus, the problem of determining whether there is a match is reduced to finding a diagonal line with slope near 1 within the scatterplot of the points in the scatter list.

There are many ways of finding the diagonal line. A preferred method starts by subtracting $m*landmark_n$ from both sides of the above equation.

$$(landmark^*_n - m^*landmark_n) = offset$$

Assuming that m is approximately 1, we arrive at

$$(landmark^*_n - landmark_n) = offset$$

The diagonal-finding problem is then reduced to finding multiple landmark pairs that cluster near the same offset value. This is accomplished easily by calculating a histogram of the resulting offset values and searching for the offset bin with the highest number of points. Since the offset must be positive if the exogenous sound signal is fully contained within the correct library sound recording, landmark pairs that result in a negative offset are excluded.

The winning offset bin of the histogram is noted for each qualifying sound_ID, and the corresponding score is the number of points in the winning bin. The sound recording in the candidate list with the highest score is chosen as the winner. The winning sound_ID is provided to an output means to signal the success of the identification.

To prevent false identification, a minimum threshold score may be used to gate the success of the identification process. If no library sound recording meets the minimum threshold then there is no identification.

Pipelined recognition

In a realtime system the sound is provided to the recognition system incrementally over time. In this case it is possible to process the data in chunks and to update the index set incrementally. Each update period the newly augmented index set is used as above to retrieve candidate library sound recordings using the searching and scanning steps above. The advantage of this approach is that if sufficient data has been collected to identify the sound recording unambiguously then the data acquisition may be terminated and the result may be announced.

Reporting the result

Once the correct sound has been identified, the result is reported. Among the result-reporting means, this may be done using a computer printout, email, SMS text messaging to a mobile phone, computer-generated voice annotation over a telephone, posting of the result to an internet account which the user can access later.